Estimation of Speech Features of Glottal Excitation by Nonlinear Prediction

Karl Schnell, Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt
Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany
schnell@iap.uni-frankfurt.de

Abstract

Analysis of speech signals can be performed with the aid of linear or nonlinear statistics using appropriate prediction algorithms. In this contribution, speech features are treated using the results of a nonlinear prediction based on Volterra series. Features are investigated representing the prediction gain by nonlinear statistics and representing individual coefficients of the nonlinear components. The features are estimated quasi continuously resulting in a feature signal. Additionally, to obtain features which are highly sensitive to segmentation shifting, an asymmetric window function is integrated into the prediction algorithm. The analyses of speech signals show that the estimated features correlate with the glottal pulses. Furthermore, the investigations show that using the first individual nonlinear coefficient as a feature is advantageous over using the prediction gain.

1. Introduction

Speech analysis is usually performed using linear models and statistics. However, nonlinear components are also contained in the speech signal [1]. The voiced excitation is caused by vibrations of the vocal folds which can be described by a nonlinear oscillator; additionally nonlinear fluid dynamics are effective. Nonlinear systems and operators, like the energy operator, can be used for speech analysis [2],[3]. In this contribution nonlinear components of the speech signal are estimated by nonlinear prediction. The nonlinear system of a Volterra series is used for the prediction. The estimation can be achieved by an adaptive algorithm like LMS or RLS [4]. Another approach for the estimation is to minimize the prediction error of individual signal segments, which can be applied to coding [5] or speech generation [6]. For speech analysis the integration of an appropriate window function can be relevant [7]. In [7] speech features based on the prediction gain are discussed. In this contribution, features of nonlinear coefficients of the predictor are proposed delivering feature signals advantageously for analysis. Additionally, a post-processing of the feature signal is carried out accentuating the regions of glottal closures.

2. Nonlinear Prediction

The nonlinear predictor based on Volterra systems estimates a signal value $x(n)$ by a linear combination of last signal values $x(n-k)$ and, additionally, by a linear combination of products of last signal values. Here, without loss of generality systems are treated with the first and second order Volterra kernels only, leading to the prediction error $e = x(n) - \hat{x}(n)$:

$$= x(n) - \sum_{k=1}^{N} h_1(k) \cdot x(n-k) - \sum_{i=1}^{M} \sum_{k=1}^{i} h'_2(i,k) \cdot x(n-i)x(n-k) . \quad (1)$$

e is the prediction error and $\hat{x}(n)$ is the estimation of $x(n)$. The coefficients h_1 represent the linear components whereas h'_2 represent the nonlinear components; h'_2 are coefficients of the second-order kernel h_2, which can be assumed symmetrically $h'_2(i,k) = h_2(i,k)$ for $i = k$ and $h'_2(i,k) = 2 \cdot h_2(i,k)$ for $i \neq k$. For speech analysis the speech signal is segmented in frames. Due to the segmentation a window function $w(n)$ is integrated into the estimation of the nonlinear prediction. If the window function is applied directly to the signal $x(n)$ the prediction error results in

$$e = w(n)x(n) - \sum_{k=1}^{N} h_1(k) \cdot \underline{w(n-k)} x(n-k) - \sum_{i=1}^{M} \sum_{k=1}^{i} h'_2(i,k) \cdot \underline{w(n-i)w(n-k)} x(n-i)x(n-k)$$

leading to different weights of the components, especially between the linear and nonlinear components. For this reason the window function has to be applied to the error $e(n)$ yielding the weighted error $e_w(n) = w(n) \cdot e(n)$. Applying to eq. (1) results in

$$e_w(n) = w(n) \cdot x(n) - \sum_{k=1}^{N} h_1(k) \cdot w(n) \cdot x(n-k) - \sum_{i=1}^{M} \sum_{k=1}^{i} h'_2(i,k) \cdot w(n) \cdot x(n-i)x(n-k) . \quad (2)$$

The predictor coefficients are determined by minimizing the weighted error

$$\sum_n e_w(n)^2 \rightarrow \min , \quad (3)$$

which is explained in the following section.

2.1. Vector based nonlinear prediction

The prediction is applied to a segment of the speech signal, so that it is convenient to describe the signals by vectors. For

that purpose the analyzed weighted signal $u(n) = w(n) \cdot x(n)$ is described by the vector

$$\boldsymbol{u} = \left(w(0) \cdot x(0), w(1) \cdot x(1), \ldots, w(K) \cdot x(K)\right)^{\mathrm{T}}$$

of length $L = K + 1$. Since the prediction error $e_w(n)$ contains last values $x(n-k)$, additionally the vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i,k}$ containing the shifted signals with fixed weights are defined by

$$\begin{aligned} \boldsymbol{u}_i &= (w(0) \cdot x(-i), w(1) \cdot x(1-i), \ldots, w(K) \cdot x(K-i))^{\mathrm{T}} \\ \boldsymbol{u}_{i,k} &= \left(w(0)x(-i)x(-k),\ w(1)x(1-i)x(1-k), \ldots\right)^{\mathrm{T}}. \end{aligned} \tag{4}$$

The estimation of the weighted signal values $u(n)$ can be described by the vector $\hat{\boldsymbol{u}}$ with

$$\hat{\boldsymbol{u}} = \sum_{i=1}^{N} h_1(i) \cdot \boldsymbol{u}_i + \sum_{i=1}^{M} \sum_{k=1}^{i} h'_2(i,k) \cdot \boldsymbol{u}_{i,k} \ . \tag{5}$$

By these definitions the prediction problem can be described by the vector equation $\boldsymbol{e}_w = \boldsymbol{u} - \hat{\boldsymbol{u}}$. Since the error depends on the order N of linear coefficients and order M of nonlinear coefficients, the error $\boldsymbol{e}_w \to \boldsymbol{e}_w^{N,M}$ is extended by the superscripts N and M:

$$\boldsymbol{e}_w^{N,M} = \boldsymbol{u} - \sum_{i=1}^{N} h_1(i) \cdot \boldsymbol{u}_i - \sum_{i=1}^{M} \sum_{k=1}^{i} h'_2(i,k) \cdot \boldsymbol{u}_{i,k} \ , \tag{6}$$

respectively

$$\boldsymbol{e}_w^{N,M} = \begin{pmatrix} e_w^{N,M}(0) \\ e_w^{N,M}(1) \\ \vdots \\ e_w^{N,M}(K) \end{pmatrix} = \begin{pmatrix} w(0)x(0) \\ w(1)x(1) \\ \vdots \\ w(K)x(K) \end{pmatrix} - h_1(1) \begin{pmatrix} w(0)x(-1) \\ w(1)x(0) \\ \vdots \\ w(K)x(K-1) \end{pmatrix} \cdots$$

$$- h'_2(1,1) \begin{pmatrix} w(0)x^2(-1) \\ w(1)x^2(0) \\ \vdots \\ w(K)x^2(K-1) \end{pmatrix} - h'_2(1,2) \begin{pmatrix} w(0)x(-1)x(-2) \\ w(1)x(0)x(-1) \\ \vdots \\ w(K)x(K-1)x(K-2) \end{pmatrix} \cdots .$$

Equation (6) represents a vector based description of eq. (2). From the equations (5) and (6) it can be seen that the optimal prediction $\hat{\boldsymbol{u}}$ is an expansion of $\boldsymbol{u}$ by the vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i,k}$. For this expansion the vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i,k}$ are transformed into an orthogonal basis $\{\boldsymbol{v}_m\}$ with the dot products $\langle \boldsymbol{v}_m, \boldsymbol{v}_k \rangle = 0$. This is performed by the Gram-Schmidt orthogonalization. Since the vectors of the basis $\{\boldsymbol{v}_m\}$ are orthogonal, the optimal coefficients b_m in description of the basis $\{\boldsymbol{v}_m\}$ can easily be obtained by

$$b_m = \langle \boldsymbol{u}, \boldsymbol{v}_m \rangle / |\boldsymbol{v}_m|^2 \ ,$$

yielding an expansion with the vectors $\boldsymbol{v}_m$. Finally the coefficients b_m of basis $\{\boldsymbol{v}_m\}$ are converted back into the original basis of $\{\boldsymbol{u}_i, \boldsymbol{u}_{i,k}\}$. The resulting coefficients minimize the Euclidean norm $|\boldsymbol{e}_w^{N,M}|$ representing a least square estimation.

Since in eqs. (2), (6) signal values outside of the frame appear, represented by negative arguments of $n-k$, the vector lengths are truncated in such a way that only values inside of the analyzed segment appear in the vectors.

3. Speech Features

The results of the nonlinear prediction can be utilized to define speech features. One approach is to consider the prediction gain by the nonlinear components. The gain can be described by the ratio between the prediction errors with and without nonlinear components, which is used in [7]. The logarithmic error ratio leads to the feature definition:

$$F_{\mathrm{gain}}^{N,M} = \log\left(\frac{|\boldsymbol{e}_w^{N,0}|}{|\boldsymbol{e}_w^{N,M}|}\right) \ .$$

Nonlinear coefficients $h'_2(i,k)$ are considered for the prediction error of the denominator, which can be seen from the superscript M. Since the nonlinear coefficients contribute only to a decrease of the prediction error, the feature $F_{\mathrm{gain}}^{N,M}$ has positive values.

Another approach for defining features is to consider individual values of the estimated predictor coefficients, especially these of the nonlinear components. Here the value of the nonlinear coefficients $h'_2(i,k)$ of the prediction of orders N and M are used leading to the feature

$$F_{i,k}^{N,M} = h'_2(i,k) \ .$$

3.1. Feature signals

The feature F is obtained from the results of the nonlinear prediction. To consider the time-dependence of the feature, the speech signal is segmented into overlapping frames analyzed individually. Applying the nonlinear prediction to each segment yields the corresponding values of the speech feature F. To measure the features quasi continuously in time, the displacement of the segments is chosen to one sample. Hence, the sequence of the feature values which are estimated from the segments represents a feature signal $F(n)$. Each value $F(n)$ is obtained from the nonlinear prediction of a segment. The estimation is influenced by the type of the window function w of the prediction algorithm. If a Hann-window is used, the feature F can be estimated smoothly in time, however, the time resolution of the feature estimation is blurred. This behaviour is caused by the shape of the window; the Hann window is insensitive to small changes of the segmentation since its values tend continuously towards zero to the left and right side. In contrast to that, an asymmetric window with a value greater one at one side is sensitive to small changes of the segmentation and can deliver a more precise time resolution. The window w_a defined by

$$\begin{aligned} w_a(k) &= 1 \quad \text{for} \quad k = 0 \ldots K/4 - 1 \\ w_a(k) &= \left(0.5\left(1 + \cos\left(\frac{\pi(k - K/4)}{K - K/4}\right)\right)\right)^2 \quad \text{for} \quad k = \tfrac{K}{4} \ldots K \end{aligned}$$

delivering a strong discontinuity between the left-side values and values outside of the window, which can be assumed as zero. The asymmetric window is shown in fig. 1.

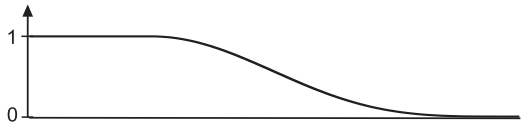

Figure 1: Asymmetric window function w_a .

4. Analysis of Speech

For the analysis of individual sounds and speech utterances, speech signals with a sampling rate of 16 kHz are investigated. The speech signals are segmented and analyzed as described in the previous section. Fig. 2 shows the estimated feature signals $F_{\text{gain}}^{16,1}(n)$ and $F_{1,1}^{16,1}(n)$ from the analysis of the vowel /a/; additionally, the original speech waveform and the LPC-residual is shown. The main impulses of the residual of fig. 2(b) indicate the abrupt glottal closures, which are denoted as the glottal closure instances (CGI). It can be seen that the feature signal has peaks correlating with those of the residual. Hence, the peaks of the feature signals indicate the CGI. The high time resolution of the feature signal results from the asymmetric window. In the case of voiced fricatives often many impulses occur in the residual, which can be seen from fig. 3 showing the analysis of the voiced fricative /z/. The LPC-residual shows a more or less unperiodic structure and especially the high incidence of pulses makes it hard to detect the impulses corresponding to the glottal closures. In contrast to that, the feature signals are more periodic and have fewer pulses. The analyses show that the feature signal $F_{1,1}^{16,1}(n)$ shows even mostly only one dominant positive pulse per period corresponding to the glottal closure; therefore the feature $F_{1,1}^{16,1}(n)$ is advantageous in comparison to the prediction gain delivering often more potential pulses per period. One reason for that is given in the following: For the prediction with order $M=1$ only the nonlinear coefficient $h'_2(1,1)$ is effective. The prediction gain depends on the absolute value of the coefficient, whereas the feature $F_{1,1}^{16,1}(n)$ preserves the information about the sign of the coefficient $h'_2(1,1)$. Analysis results show that the glottal closures cause impulses with positive sign. Other regions of the feature signal show also impulses or bulges, however, for $F_{1,1}^{16,1}(n)$ they have usually negative sign. Hence, by the feature signal $F_{1,1}^{16,1}(n)$ the impulses of glottal closure can be separated from the other regions with the aid of the sign of $h'_2(1,1)$. In comparison to that the feature $F_{\text{gain}}^{16,1}(n)$ cannot distinguish since the information of the sign is lost in the prediction gain.

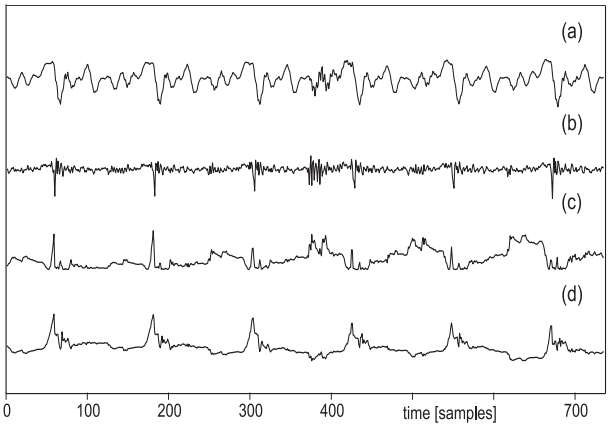

Figure 2: Analysis of the vowel /a/: (a) analyzed speech signal, (b) corresponding LPC-residual, (c) feature signal of prediction gain $F_{\text{gain}}^{16,1}(n)$, (d) feature signal $F_{1,1}^{16,1}(n)$.

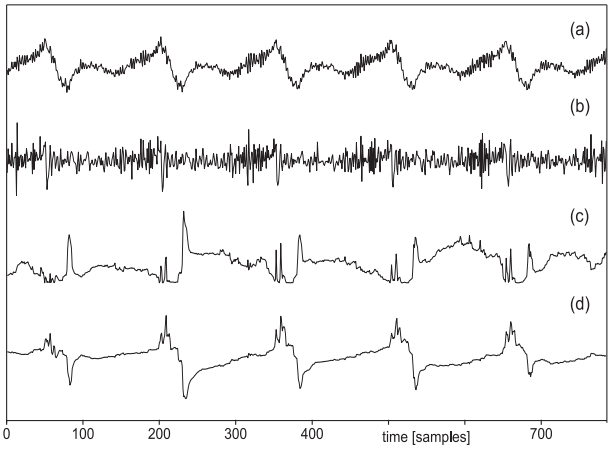

Figure 3: Analysis of the voiced fricative /z/: (a) analyzed speech, (b) corresponding LPC-residual, (c) feature signal of prediction gain $F_{\text{gain}}^{16,1}(n)$, (d) feature signal $F_{1,1}^{16,1}(n)$.

Overall, the analyses show that especially the feature signal $F_{1,1}^{16,1}(n)$ is suitable for detection of regions of glottal closures not only for stationary speech signals, but also for speech utterances. A post-processing of the feature signal is useful to mark the regions of glottal closures. At first, fluctuations of the mean value differing from zero should be compensated; additionally, the power of the feature signal should be balanced achieving a constant envelope of the amplitude. Therefor a short-time estimation of the mean of the feature signal is subtracted to each feature value. After that, each feature value is divided by a short-time estimation of the power of the feature signal resulting in the modified feature signal $\tilde{F}_{1,1}^{16,1}(n)$.

Figure 4 shows the analysis of the German word [nUl]. The curves 4(c) and (d) show the initial feature and the modified feature signal $\tilde{F}_{1,1}^{16,1}(n)$; variations of the mean and the power of the feature signal are balanced. After that the modified

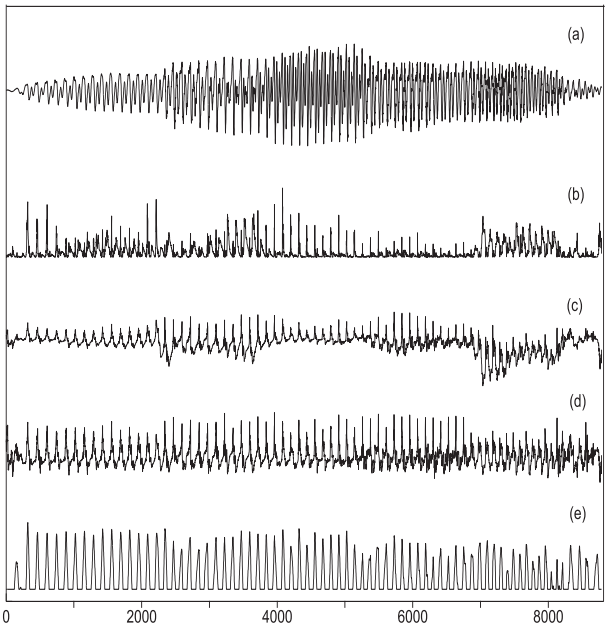

Figure 4: Analysis of word [nUl]: (a) analyzed speech signal, (b) feature signal of prediction gain $F_{\text{gain}}^{16,1}(n)$, (c) feature signal $F_{1,1}^{16,1}(n)$, (d) processed feature $\tilde{F}_{1,1}^{16,1}(n)$, (e) derived feature signal $f'(n)$.

feature signal is convolved by a finite pattern-signal g depicted in fig. 5 which has a pointed shape resulting in the signal

$$f(n) = \tilde{F}_{1,1}^{16,1}(n) * g(n) ;$$

the mean value of the signal g is zero. The convolution implies dot products with time-shifted segments. If the segment is similar to the pointed shape, a high value results.

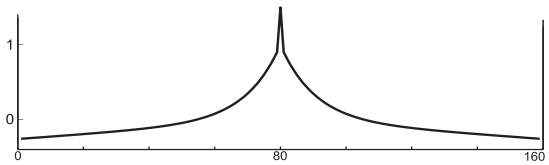

Figure 5: Pattern-signal *g*.

Since only positive correlations with the pointed shape are of interest, negative values of *f* are set to zero by

$$f'(n) = (f(n) + \text{sgn}(f(n)) \cdot f(n))/2 .$$

The curve 4(e) shows the derived feature signal $f'(n)$ representing the positive values of the convolution results of the modified feature signal of the utterance [nUl]. The peaks indicate regions of glottal pulses.

In figure 6 the analysis result for the utterance [vaIma] of the German word “Weimar” is shown. It can be seen that the corresponding feature signal $f'(n)$ represents a sequence of pulses, which is disturbed only occasionally by artifacts.

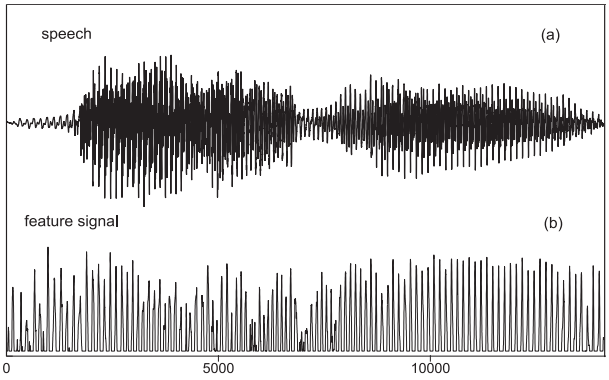

Figure 6: Analysis of word [vaIma]: (a) analyzed speech signal, (b) derived feature signal $f'(n)$.

5. Conclusions

Speech features based on nonlinear prediction are proposed and discussed for the analysis of speech. The features are correlated with the voiced excitation and especially with the glottal pulses. For analysis of real speech, one important feature of estimation algorithms is their robustness. Concerning this, features based on the first nonlinear prediction coefficient have been proven advantageously in comparison to the prediction gain. The deciding reason for this fact is that the informational content of the sign of the coefficient is useful. By the use of that feature signal with an additional post-processing the algorithm is applicable to analyse real speech.

6. References

[1] M. Faundez et al., "Nonlinear Speech Processing: Overview and Applications", in *Int. J. Control Intelligent Syst., vol. 30, no. 1*, pp. 1–10, 2002.

[2] P. Maragos, T. Quatieri, and J. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators", in *Proc. ICASSP’91*, 1991, pp. 421-424.

[3] L. Atlas and J. Fang, "Quadratic Detectors for General Nonlinear Analysis of Speech", in *Proc. ICASSP’92, vol. II*, 1992, pp. 9–12.

[4] E. Mumolu, A. Carini, and D. Francescato, "ADPCM With Non Linear Predictors", in *Proc. EUSIPCO’94*, 1994, pp. 387–390.

[5] J. Thyssen, H. Nielsen, and S. D. Hansen, "Non-linear Short-term Prediction in Speech Coding", in *Proc. ICASSP’94, vol. I*, 1994 pp. 185–188.

[6] K. Schnell and A. Lacroix, "Modeling Fluctuations of Voiced Excitation for Speech Generation Based on Recursive Volterra Systems", contribution in *Nonlinear Analyses and Algorithms for Speech Processing – NOLISP’05, LNAI Vol. 3817*, pp. 338-347, Springer 2005.

[7] K. Schnell and A. Lacroix, "Weighted Nonlinear Prediction Based on Volterra Series for Speech Analysis", in *Proc. EUSIPCO’06*, Florence 2006.